HEP Computing in Switzerland

Christoph Grab (ETH)

Head of CHIPP Computing Group

R-ECFA visit, April 1, 2016

Status of WLCG Tier-2 and Tier-3 computing resources in Switzerland

Worldwide LHC Computing : WLCG

Expts

Tier 0

store raw data,
reconstruct + distribute

Tier 1

store +redistribute data
reconstruct, analyse

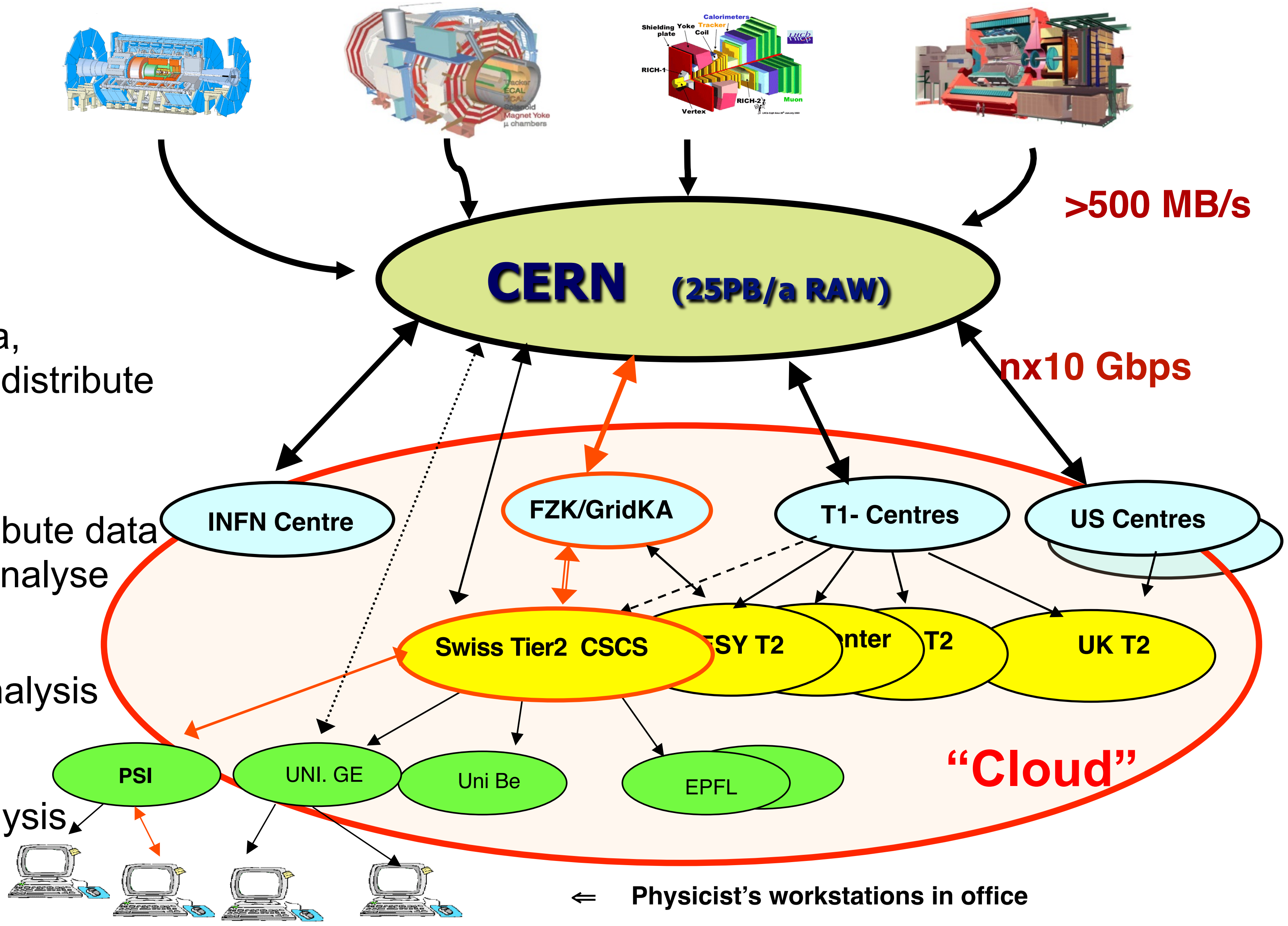
Tier 2

simulation, analysis

Tier 3

end-user analysis

Tier 4



>500 MB/s

CERN (25PB/a RAW)

nx10 Gbps

INFN Centre

FZK/GridKA

T1- Centres

US Centres

Swiss Tier2 CSCS

...SY T2

...nter

T2

UK T2

PSI

UNI. GE

Uni Be

EPFL

“Cloud”

⇐ Physicist’s workstations in office

Overview Swiss LHC Computing Resources

- **Switzerland operates a single Tier-2 Regional Centre at CSCS**
  - Maintain our own dedicated compute-cluster integrated into "WLCG" .
  - Switzerland is committed as full member to contribute resources; signed MoU
- **Tier-2 operated by CSCS, serves all 3 experiments: ATLAS,CMS, LHCb**
  - Collaboration agreement for operation of T2 between CHIPP and CSCS/ETHZ (2007-2018 with additional ETHZ funding secured)
  - Presently: 90% of available resources provided to WLCG and exploited centrally by experiments; other ~10% reserved for Swiss users only

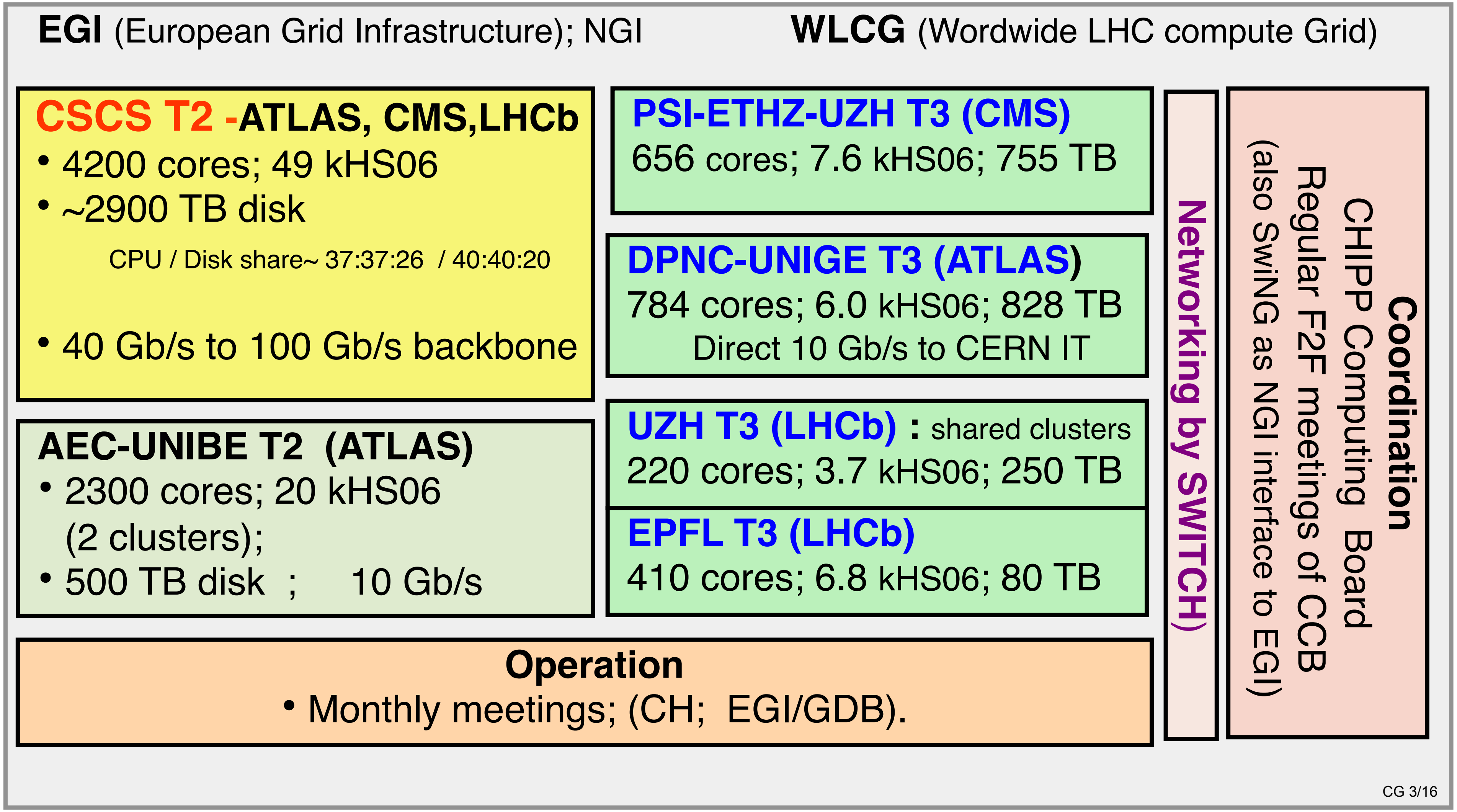
CSCS
Centro Svizzero di Calcolo Scientifico
Swiss National Supercomputing Centre

Swiss Tier-2 Phoenix cluster at Lugano

- **CSCS Tier-2 supplemented by ATLAS only resources** at AEC-UNIBE
- **Complemented by local Tier-3 clusters** at PSI, UBe+UGe, UZH+EFL

CHiPP Swiss Institute of Particle Physics

Overview Swiss LHC Computing Resources



EGI (European Grid Infrastructure); NGI

WLCG (Wordwide LHC compute Grid)

CSCS T2 -ATLAS, CMS,LHCb
- 4200 cores; 49 kHS06
- ~2900 TB disk

CPU / Disk share~ 37:37:26 / 40:40:20

- 40 Gb/s to 100 Gb/s backbone

AEC-UNIBE T2 (ATLAS)
- 2300 cores; 20 kHS06 (2 clusters);
- 500 TB disk ; 10 Gb/s

PSI-ETHZ-UZH T3 (CMS)
656 cores; 7.6 kHS06; 755 TB

DPNC-UNIGE T3 (ATLAS)
784 cores; 6.0 kHS06; 828 TB
Direct 10 Gb/s to CERN IT

UZH T3 (LHCb) : shared clusters
220 cores; 3.7 kHS06; 250 TB

EPFL T3 (LHCb)
410 cores; 6.8 kHS06; 80 TB

Operation
- Monthly meetings; (CH; EGI/GDB).

Networking by SWITCH)

Coordination
CHIPP Computing Board
Regular F2F meetings of CCB
(also SwiNG as NGI interface to EGI)

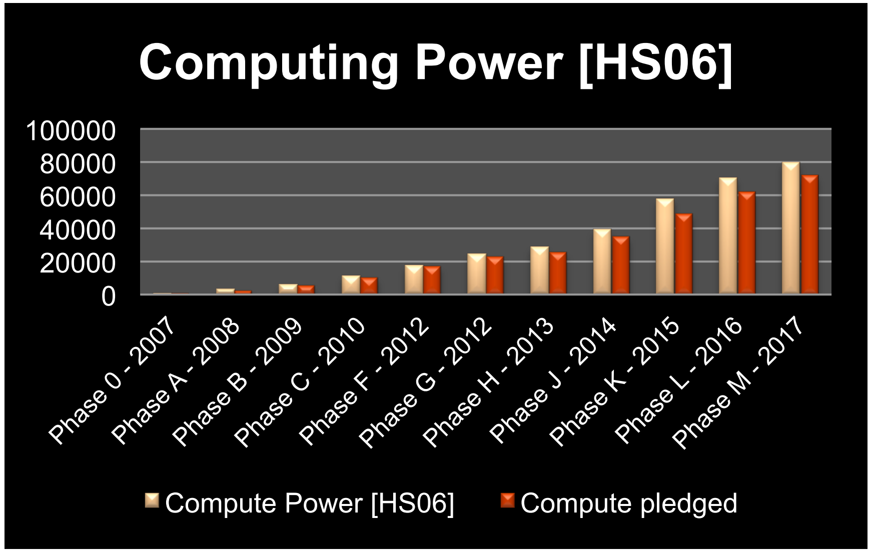
CG 3/16

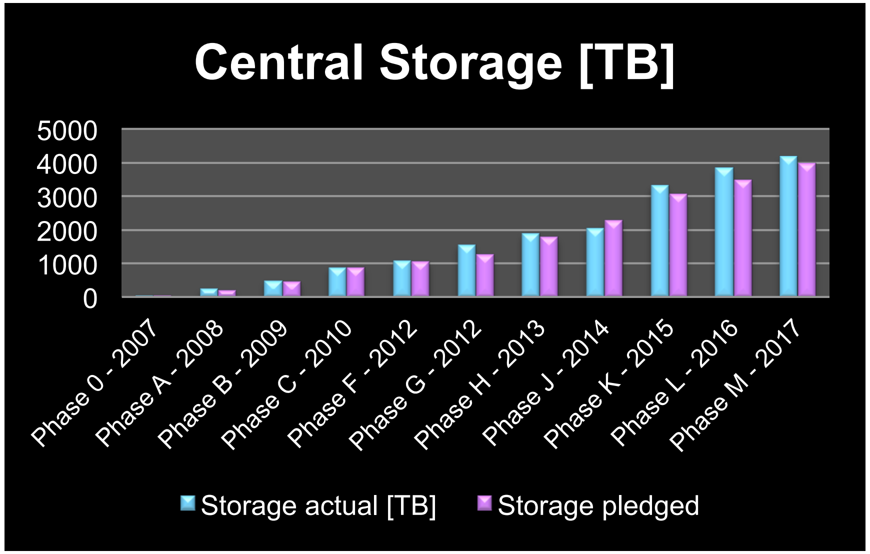
Note: sum of Tier-3 resources [25 kHS06; 1.5 PB]
equals ~ 2/3 of Tier-2 resources (except ATLAS)

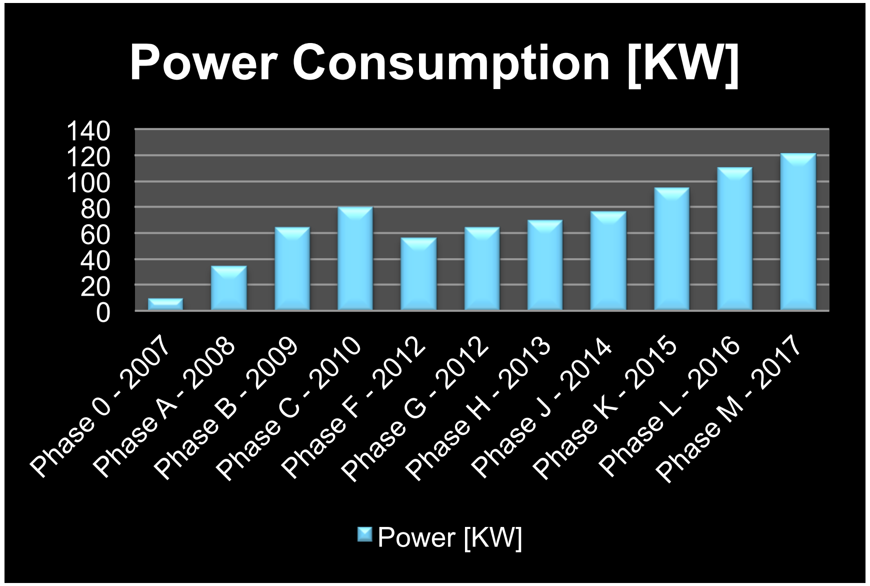
Swiss Tier-2 Resources Evolution

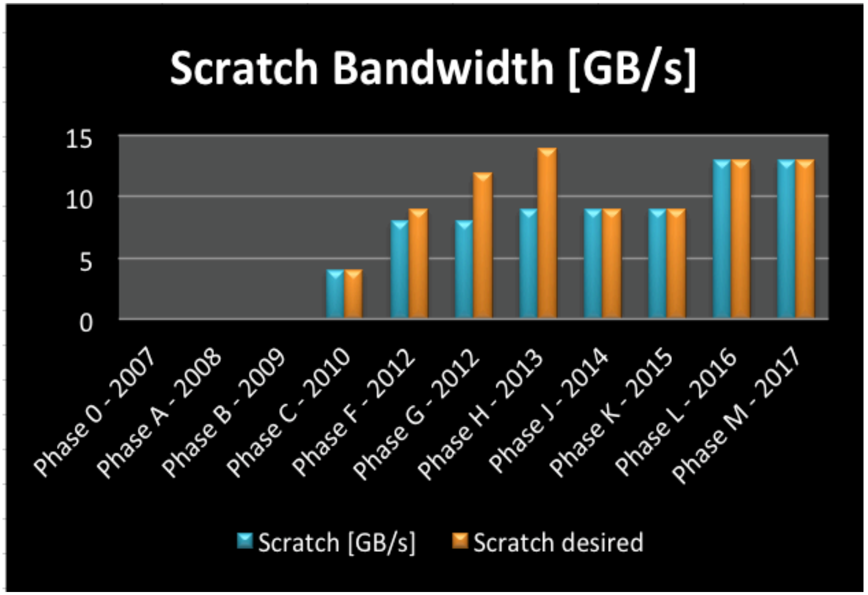
Evolution for 2007 – 2017

(phase K:= installed in 2015; meet pledges 1.4.2016)

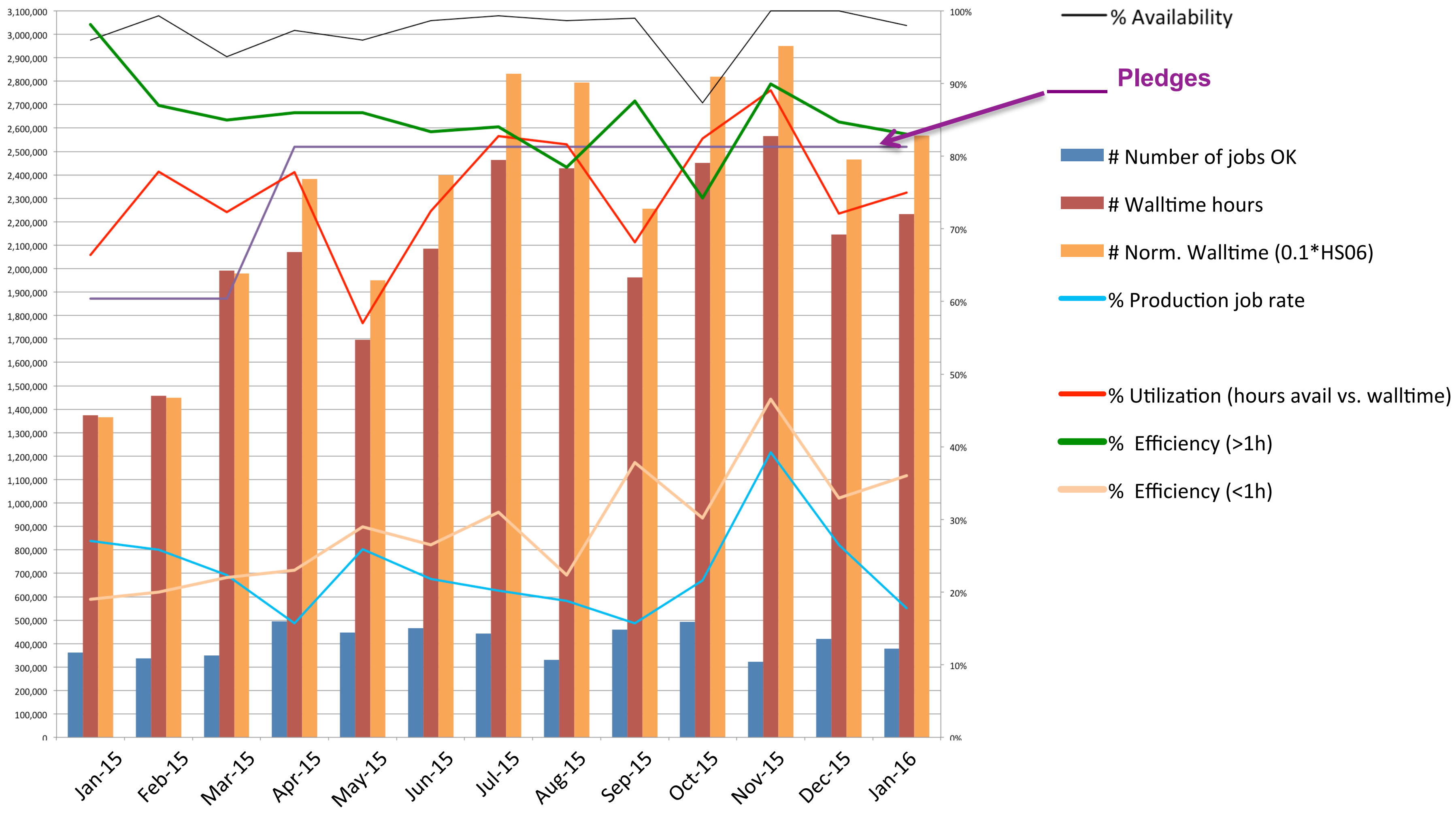








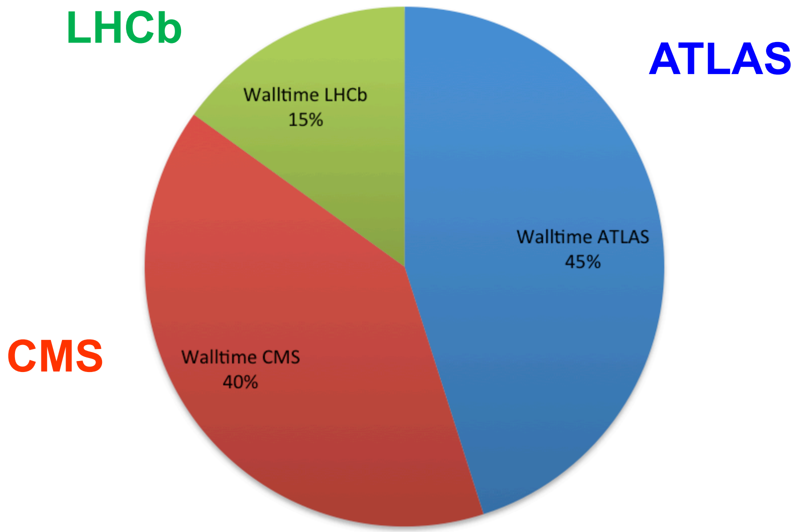
Cluster CPU statistics (2015-2016)



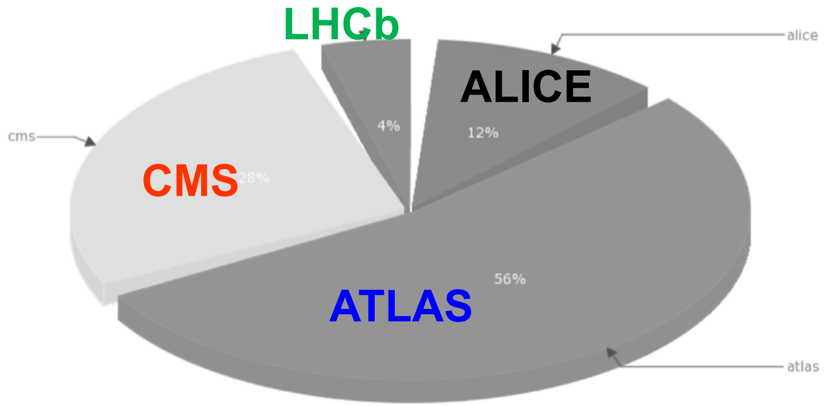
Overall high **availability** (>95%) and efficiency typ. 85-95% achieved !
Utilization at around 80 %. Pledges (walltime h) are met.

Resource distribution @CSCS 1.2015–1.2016

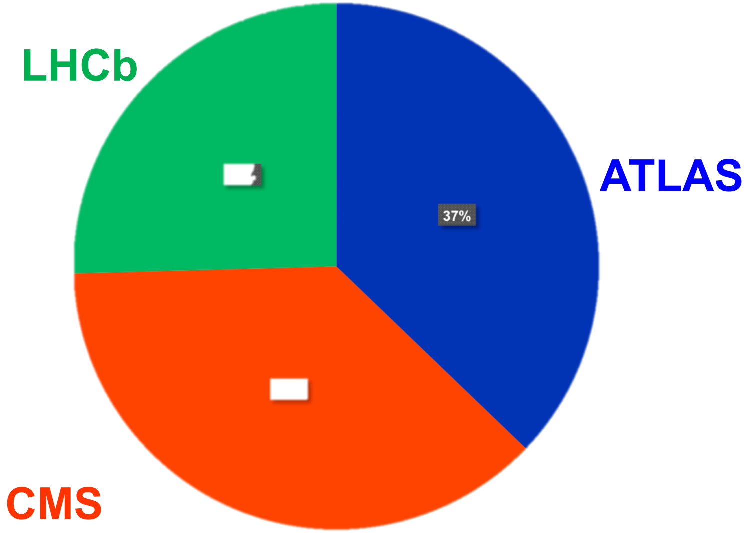
Walltime CPU usage 1.2015-1. 2016



Compare to worldwide T2 usage



Storage usage on 3.2016



Resource ratios at CSCS:
ATLAS:CMS:LHCb

- ➢ CSCS fairshare ratio 40:40:20
- ➢ effective CPU usage: 45:40:15
- ➢ CSCS disk ratio: 37:37:26

Comments on Resources

- **HW investments at CSCS (**replacements and additions) are based on C-RRB recommendations of a **"flat budget"**. Funded by FLARE/SNF Provides typically 15-20% increase of resource "power" per year.

- **Personnel for operation** :
  - 1.5 FTE to support Tier-2 operation at CSCS, covered by SNF/FLARE
  - 1 additional FTE covered by ETH internal funds
  - Additional ~0.4 FTE per experiment as user- and experiment-specific software support, covered by institutes
  - Overall management and coordination tasks covered by ETH

- **Other resource items T2 and T3**
  - Recurring power/infrastructure costs at CSCC are carried by ETH
  - Tier-3 hardware costs covered by institutes
  - specific Tier-3 manpower covered by institutes, partly by SNF

Swiss Tier-3 resources

Swiss Tier-3 resources are undispensible tools
and exist in quite different “flavours” for :

- ATLAS: each at UBern and at UGe
- CMS: common T3 for ETHZ, UZH, PSI at PSI
- LHCb: each at UZH and EPFL.

- Their capacity sum up to ~50% and 70% of CPU and storage of Tier-2 (at CSCS w/out AEC) .

ATLAS Tier-2/3 at Uni Bern

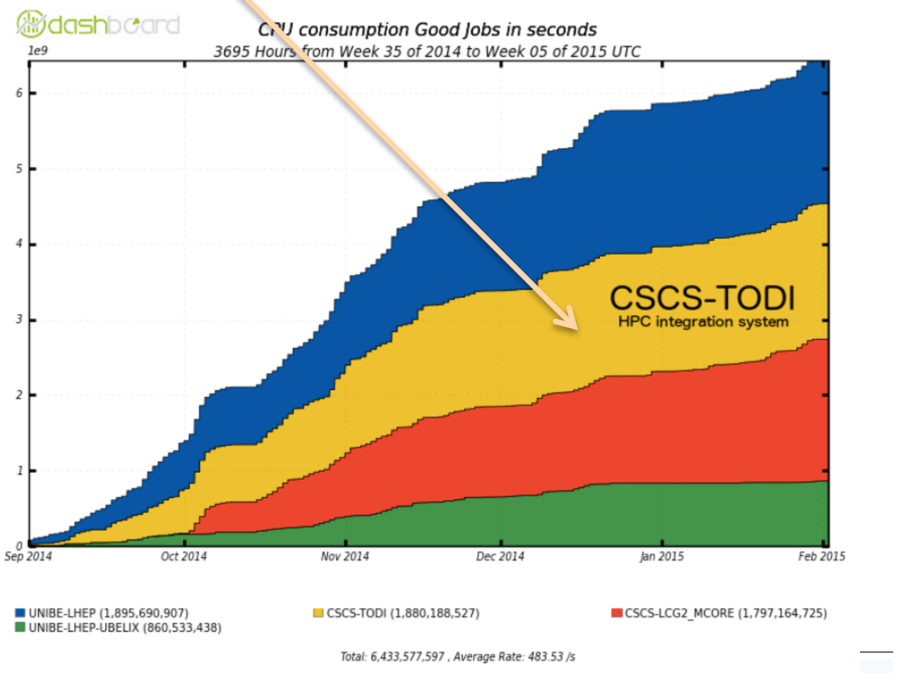
- Use three clusters (one shared university cluster and two AEC clusters) → ~2700 cores and 500 TB disk
  - Pledged for 2015: 10 kHS06, 350 TB
  - Full ATLAS MC production and analysis
  - Serve about 20 ATLAS users at AED

} **Full Tier-2 functionality for ATLAS**

- Run ATLAS MC production successfully on CSCS Cray; ATLAS is ready to move MC production to HPC at CSCS → may serve ATLAS' future Tier operation model with nuclei and satellite centres.

- AEC-Bern also serves T2K and Microboone VOs.

- Explore usage of other free resources (e.g. Switch-engines ...)



CMS Tier-3 common for ETHZ, PSI, UZH

- **A standard linux cluster located at PSI,** HW financed by the institutes; 1 FTE by SNF. Power and infrastructure by PSI.
- **Operates the full CMS software framework,** (available via /CVMFS; but no ARC-CE). **Allows crab job submission to GRID, and data stage-out to local storage.**

CPU: 650 cores; 7.6 kHS06
Grid Storage: 755 TB

6 VMS for SGE, MySQL, BDII, dCache, PostgreSQL, Ganglia, LDAP, Nagios, CMS-frontier, PhEDEx, CVMFS,...

~ 10 power users;
~ 30 total users

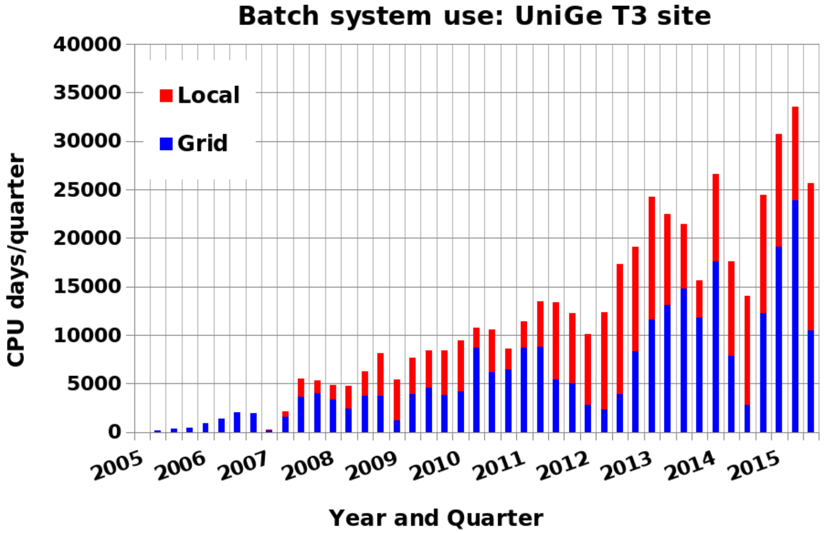
35 WNs	464 cores	6200 HS06
6 UIs	192 cores	1446 HS06
		7646 HS06

SUN x4500	4*15 TB
SUN x4540	5*31 TB
SGI IS5500	270 TB
NetApp E5400	270 TB
	~215+540 TB

The Geneva T3 resources

- Standard batch system: 784 CPU cores (5990 HS06)
  - 656 in batch, 96 login, 32 Windows
- Storage system: 828 TB for different communities
  - 474 TB in a grid Storage Element (DPM); 354 TB in NFS:
    - ATLAS (31%), neutrino (5%), AMS (29.5%), IceCube (0.5%), DAMPE (34%)
- 10 Gb/s direct to CERN & Swiss academic network
- ATLAS GRID Services: ARC-CE, DPM SE, BDII,..
  run standard ATLAS grid jobs

~16 normal,
8 power users
(in 2016)



UZH – LHCb Tier3 Resources

Status

LHCb Zürich maintains a local simulation and analysis cluster; administrated by institute.

Cluster is part of the LHCb DIRAC framework (not WLCG),
(run LHCb Grid jobs on idle CPUs)

Hardware

- 220 CPU cores (ca. 3700 HS06)
- 250 TB disk space

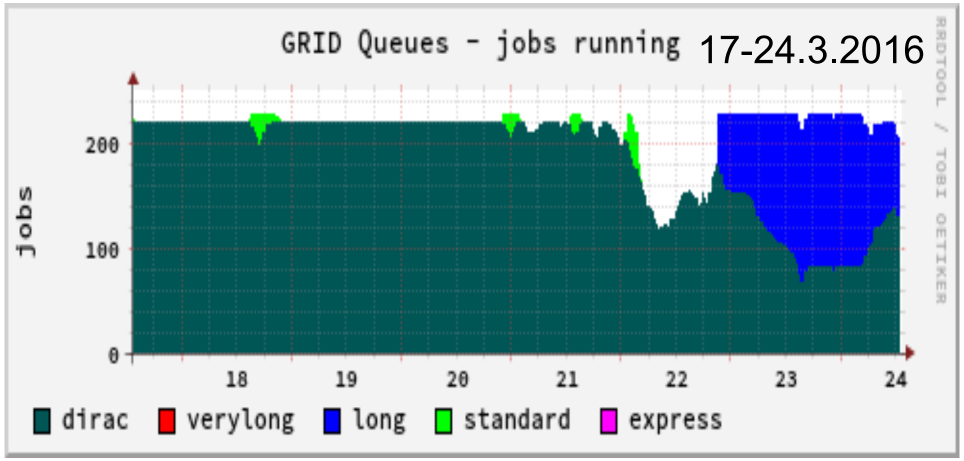
Development

- Started to use the UZH ScienceCloud, an OpenStack multi-purpose compute and storage infrastructure at University.
  Currently ~40% of the CPU power is delivered by the ScienceCloud.

Usage

Mostly dirac LCG jobs
Others: local user jobs

4 power users
4 normal users



17-24.3.2016

LHCb Tier-3 at EPFL

EPFL operates a basic linux cluster for local LHCb analysis (acquired 2014).

Standard batch system

- Use 25 nodes (16 x 2.6 GHz CPU)   CPU ~ 6.8 kHS06

Storage:

- Use simple 80 TB disk based file-system.

Analysis:

- Approximately 30 users.
- LHCb software is installed via the CernVM-FS, cached in /cvmfs..
- Cluster used for ganga job submission, or local analysis. Not a DIRAC site.

System administration:

- Faculty support for hardware.
- SCITAS (SCientific IT and Application Support) HPC support for installation of LHCb applications.

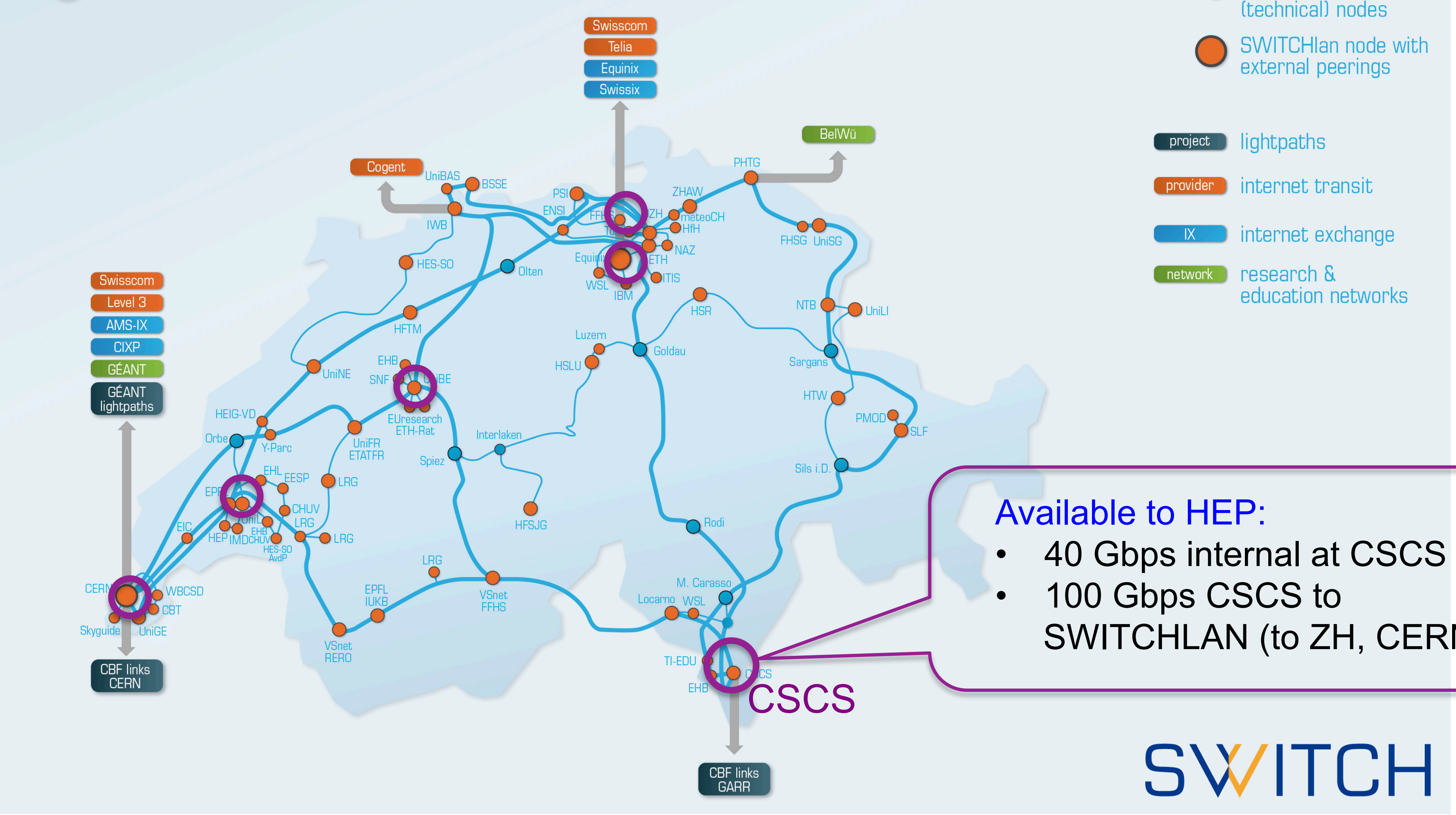
Network in Switzerland

Swiss National Network

SWITCHlan Backbone serves us/HEP well

Dec. 2015

○ Tier 2/3



CSCS

Available to HEP:

- 40 Gbps internal at CSCS
- 100 Gbps CSCS to SWITCHLAN (to ZH, CERN)

SWITCH

Swiss EGI.ch Membership status

- EGI.eu membership is a (formal) requirement for WLCG participation due to dependence on EGI paid services.
  - Switzerland is a full member since 2010. SwiNG is the formal member association with mandate from SERI (state secretariat).
  - Swiss participation is currently partly federally funded (swissuniversities). Situation beyond 2016 to be clarified.
  - AEC-LHEP University of Bern represents at European level
  - Swiss production sites integrated in EGI infrastructure are CSCS, PSI, UNIBE, UNIGE
  - Central operation and national grid Certificate Authority provided presently through AEC

- CH participated in FP7 EGI-InSPIRE via SWITCH/SwiNG/CHIPP.
- CH participates in H2020 EGI-Engage (2015-2017) via SwiNG/FMI.
- EGI.eu participates in several other projects with CH partners.

Efforts towards a Future Model of improved resource sharing

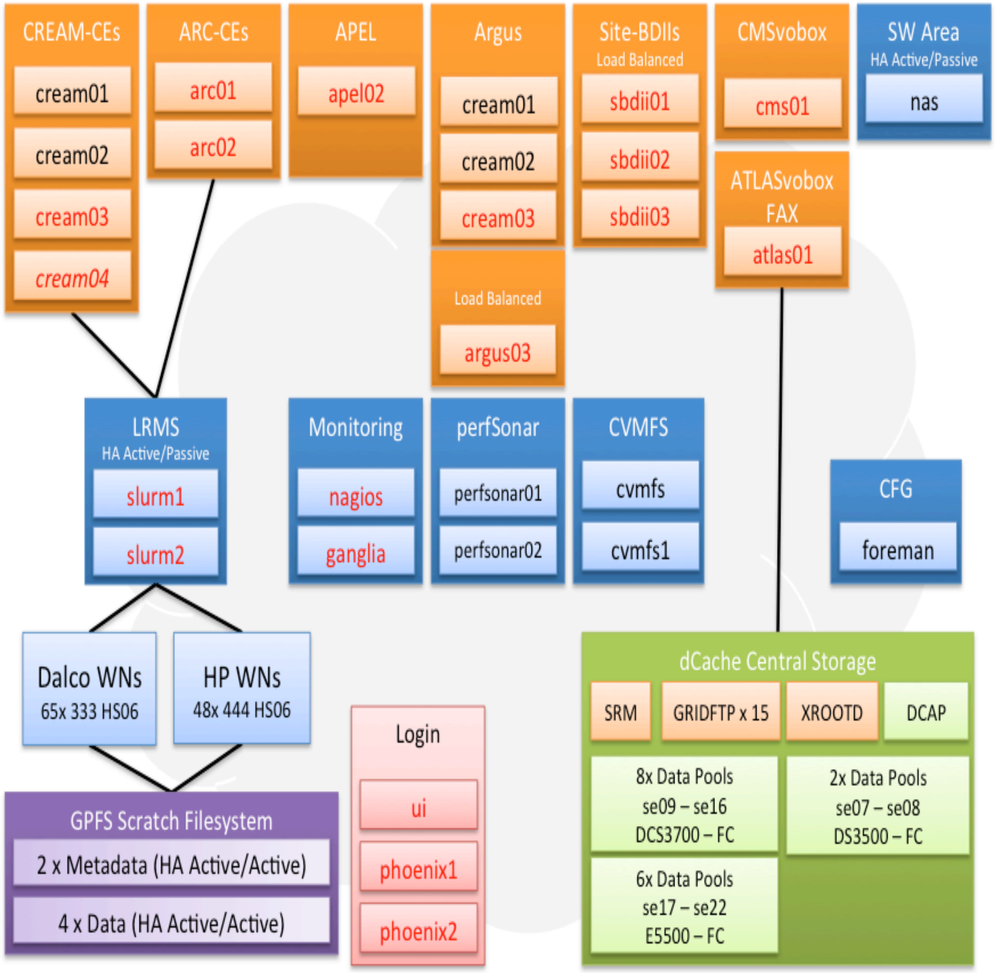
"LHConCray at CSCS"

LHConCray at CSCS – present situation

- CHiPP currently operates its own dedicated Tier-2 hardware cluster at CSCS. This requires:
  - Maintain multiple middleware interfaces (compute, storage, info)
  - All tailored specifically for CHiPP
  - System/Interfaces at CSCS, VO representatives outside

Although efficiency was increased over years by sharing resources between all 3 Vos,

→ It can be much improved by sharing resources with many other communities

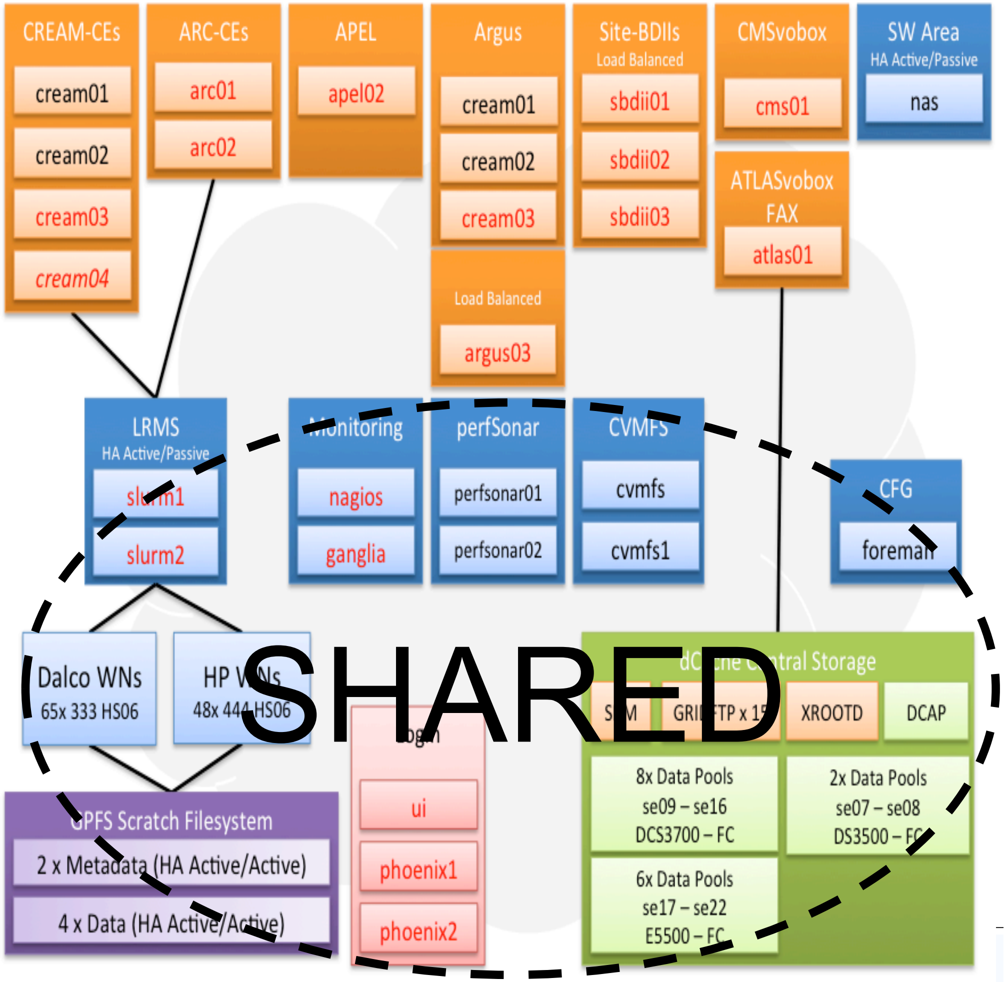


LHConCray at CSCS – goal

- Goal is to share resources and efforts with other communities.
- In our specific case this means profit from the shared HPC Systems at CSCS (with >6500 nodes and >10 PB of storage) while keeping the interfaces to the Grid World WLCG

Project requires:

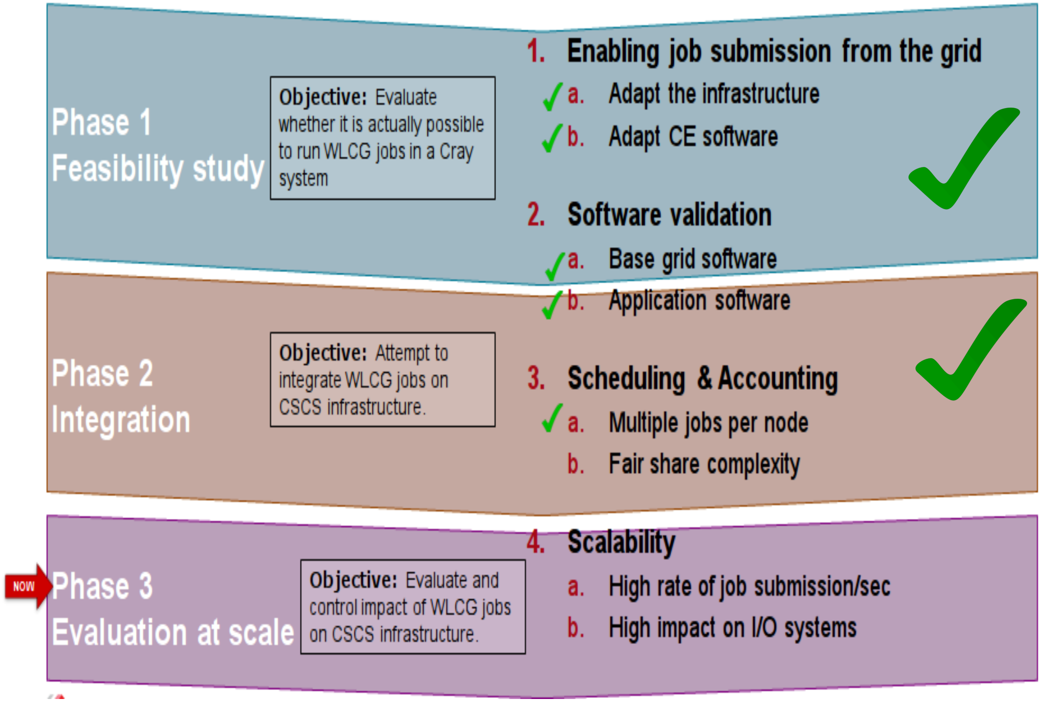
- Porting different workflows (VO Job factories and such) into the shared systems
- Render Grid Middleware CRAY-enabled
- Involve the whole Grid community
- shifting part of the resources currently spent on the Tier2 (CSCS and VO-representatives)
- Eventually decommission the Phoenix Tier2 cluster



LHConCray at CSCS – status

- *Objective: Run MonteCarlo production jobs on the CSCS Cray shared*
  → met already for ATLAS; CMS and LHCb in progress

- Project began as a proof-of-concept in early 2015 and has passed all obstacles until now
- Currently evaluating at scale with very good results so far
- Already adopted by the VO central factories for most workflows
- WILL NEED a sustainable investment model, and support by funding agencies (SNF/ FLARE) !



NOW

Phase 1
Feasibility study

Objective: Evaluate whether it is actually possible to run WLCG jobs in a Cray system

1. Enabling job submission from the grid
   - ✓ a. Adapt the infrastructure
   - ✓ b. Adapt CE software
2. Software validation
   - ✓ a. Base grid software
   - ✓ b. Application software

✓

Phase 2
Integration

Objective: Attempt to integrate WLCG jobs on CSCS infrastructure.

3. Scheduling & Accounting
   - ✓ a. Multiple jobs per node
   - b. Fair share complexity

✓

Phase 3
Evaluation at scale

Objective: Evaluate and control impact of WLCG jobs on CSCS infrastructure.

4. Scalability
   - a. High rate of job submission/sec
   - b. High impact on I/O systems

Profits:

✓ Broaden availability of resources worldwide far beyond present technologies
✓ Leverage from economy of scales when procuring hardware
✓ Reduce hardware-related operational costs (among other)
✓ Cooperate and involve other communities (HPC)...

Activities by the Neutrino Community in view of large Data Handling

Neutrino Community
WA105: CERN EHN1 test beam extension

- **Upcoming needs for computing resources by neutrino community:**
  - ✓ Online computing farms
  - ✓ High-volume data storage
  - ✓ Data access world wide

- **DUNE/WA105 Offline Computing/Analysis**.
  - High Level Data Flow: common development DUNE + CERN IT / FNAL SCD
  - Local online computing farm stores raw data from DAQ
    - 1 PB online farm isolates DAQ from CERN EOS storage
    - ≈400 CPU cores perform online event filtering, data reduction
    - Total data volume estimated to some 2.4 PB/yr
    - Beam rate = 100Hz, data flow = 15 GB/s, installing 20 Gb/s link from CERN EHN1 to IT computing centre

DUNE DEEP UNDERGROUND NEUTRINO EXPERIMENT

- **Distributed analysis model**: Data will be distributed further (CERN→FNAL→Univ/Labs) with frequent access to Raw Data in the initial phase of the experiment → requires network bandwidth

(info by A.Rubbia)

CHIPP Computing Board

Coordinates the tier-2 and tier-3 activities

includes representatives of all institutions and experiments,
CSCS, and tier-3 experts

T.Golling, Luis M.Ruiz (UNI Ge)
S.Haug, G.Sciacca (UNI Bern)

C.Grab (ETHZ) chair CCB
D.Feichtinger (PSI) vice-chair CCB
J.Pata (ETHZ), F.Martinelli (PSI)

R.Bernet (UNIZH)
A.Bay, M.Tobin (EPFL)

P.Fernandez, M.Gila, M.Ricciardi,
M. De Lorenzi (CSCS)

Thank you …

Backup slides

Comments on the national IT Landscape for Swiss Academia

Inititiatives by SUK - Swissuniversities

Swissuniversities - Initiatives

SUK-Programm 2013-2016 P-2 «Wissenschaftliche Information: Zugang, Verarbeitung und Speicherung»
"Scientific information: Access, processing and safeguarding”.

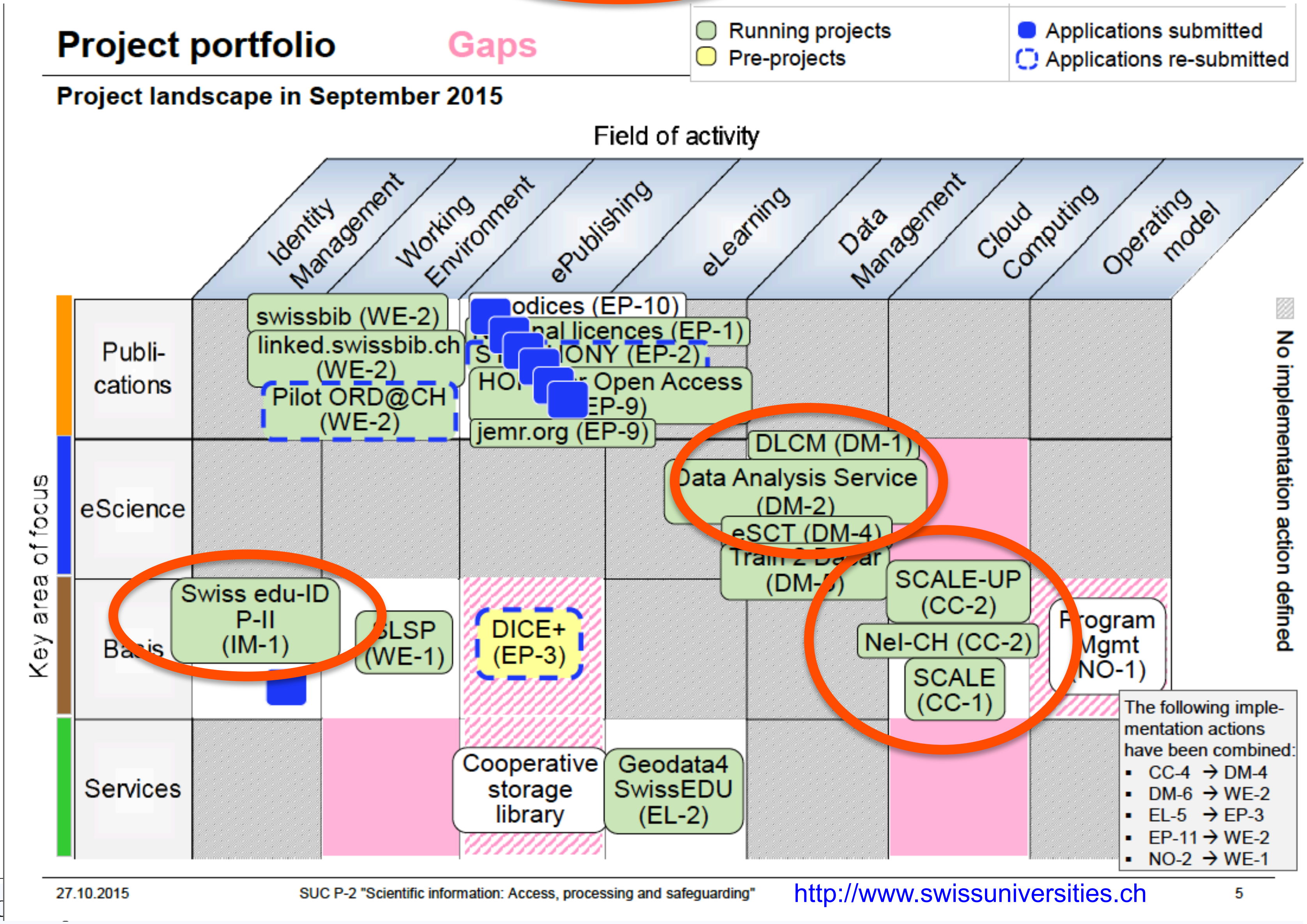
SUC P-2 national initiative for academia with funding programm (2 annual calls since 2013; total of 45 MCHF). *Has impact on HEP computing – we can profit*.

Mandat:
“SUK P-2 fördert die Bündelung und Entwicklung der heute verteilten Anstrengungen der Hochschulen für die Bereitstellung und Verarbeitung von wissenschaftlicher Information. Zur Stärkung der Schweizer Wissenschaft im internationalen Wettbewerb soll eine Neuordnung etabliert werden, die Forschenden, Lehrenden und Lernenden ein umfangreiches Grundangebot an digitalen Inhalten von wissenschaftlicher Relevanz und optimale Werkzeuge für deren Verarbeitung zur Verfügung stellt. Durch gezielte Förderung initiiert und steuert P-2 den Aufbau dieses Angebots und sorgt für einen nachhaltigen Betrieb.”

http://www.swissuniversities.ch/en/organisation/projekte-und-programme/

Swissuniversities - Initiatives

Various projects have direct **links to HEP**



Project portfolio Gaps

🟩 Running projects
🟨 Pre-projects
🟦 Applications submitted
⬚ Applications re-submitted

Project landscape in September 2015

Field of activity

The following implementation actions have been combined:
- CC-4 → DM-4
- DM-6 → WE-2
- EL-5 → EP-3
- EP-11 → WE-2
- NO-2 → WE-1

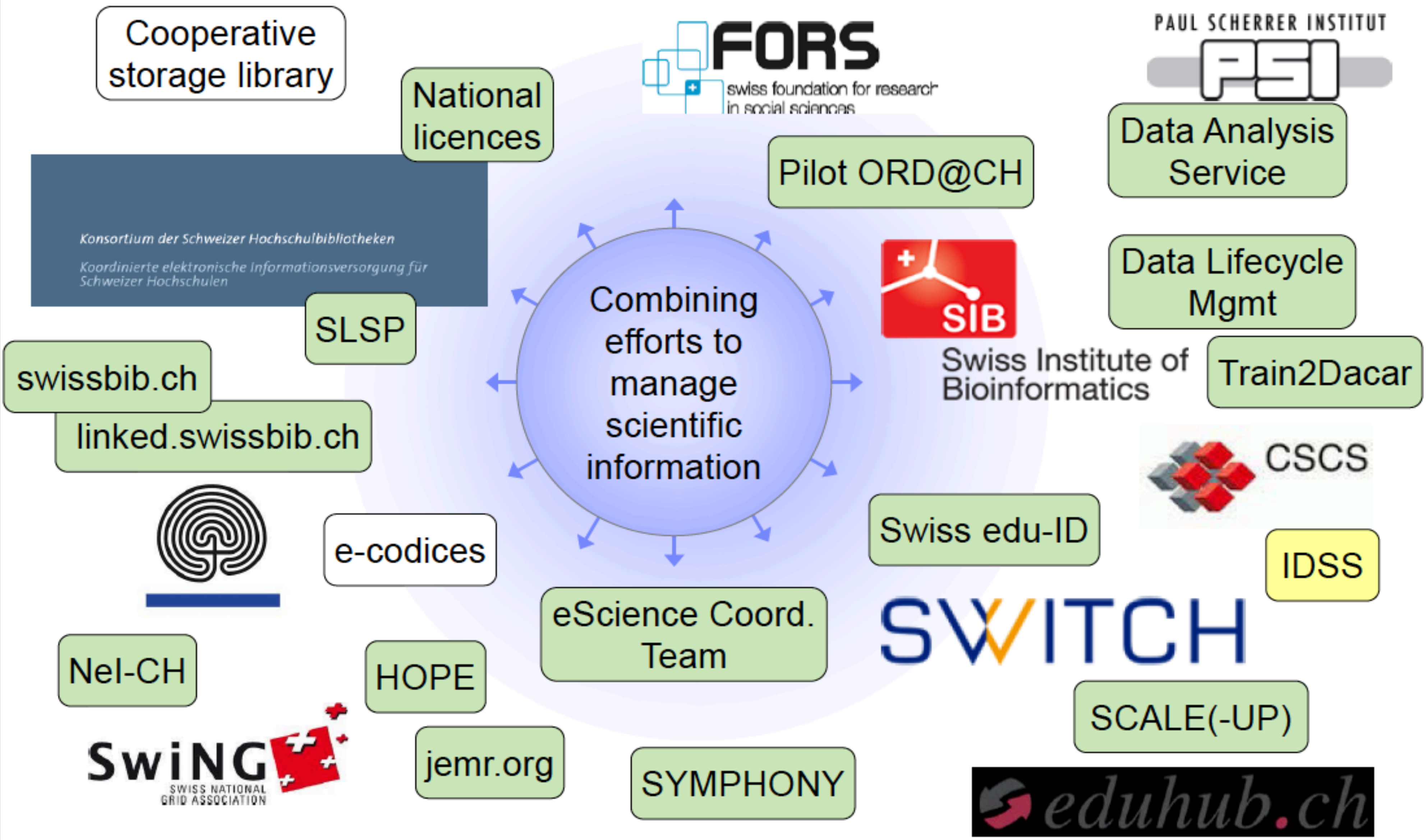
27.10.2015 SUC P-2 "Scientific information: Access, processing and safeguarding" 5

http://www.swissuniversities.ch

Swiss Institute of Particle Physics

Swissuniversities – National Initiatives

A hub of shared services



Combining efforts to manage scientific information

Cooperative storage library

National licences

FORS swiss foundation for research in social sciences

PAUL SCHERRER INSTITUT PSI

Data Analysis Service

Pilot ORD@CH

Konsortium der Schweizer Hochschulbibliotheken

Koordinierte elektronische Informationsversorgung für Schweizer Hochschulen

SLSP

swissbib.ch

linked.swissbib.ch

SIB Swiss Institute of Bioinformatics

Data Lifecycle Mgmt

Train2Dacar

CSCS

e-codices

Swiss edu-ID

IDSS

SWITCH

eScience Coord. Team

Nel-CH

HOPE

SCALE(-UP)

SwiNG SWISS NATIONAL GRID ASSOCIATION

jemr.org

SYMPHONY

eduhub.ch

Overview on CSCS Resources and services

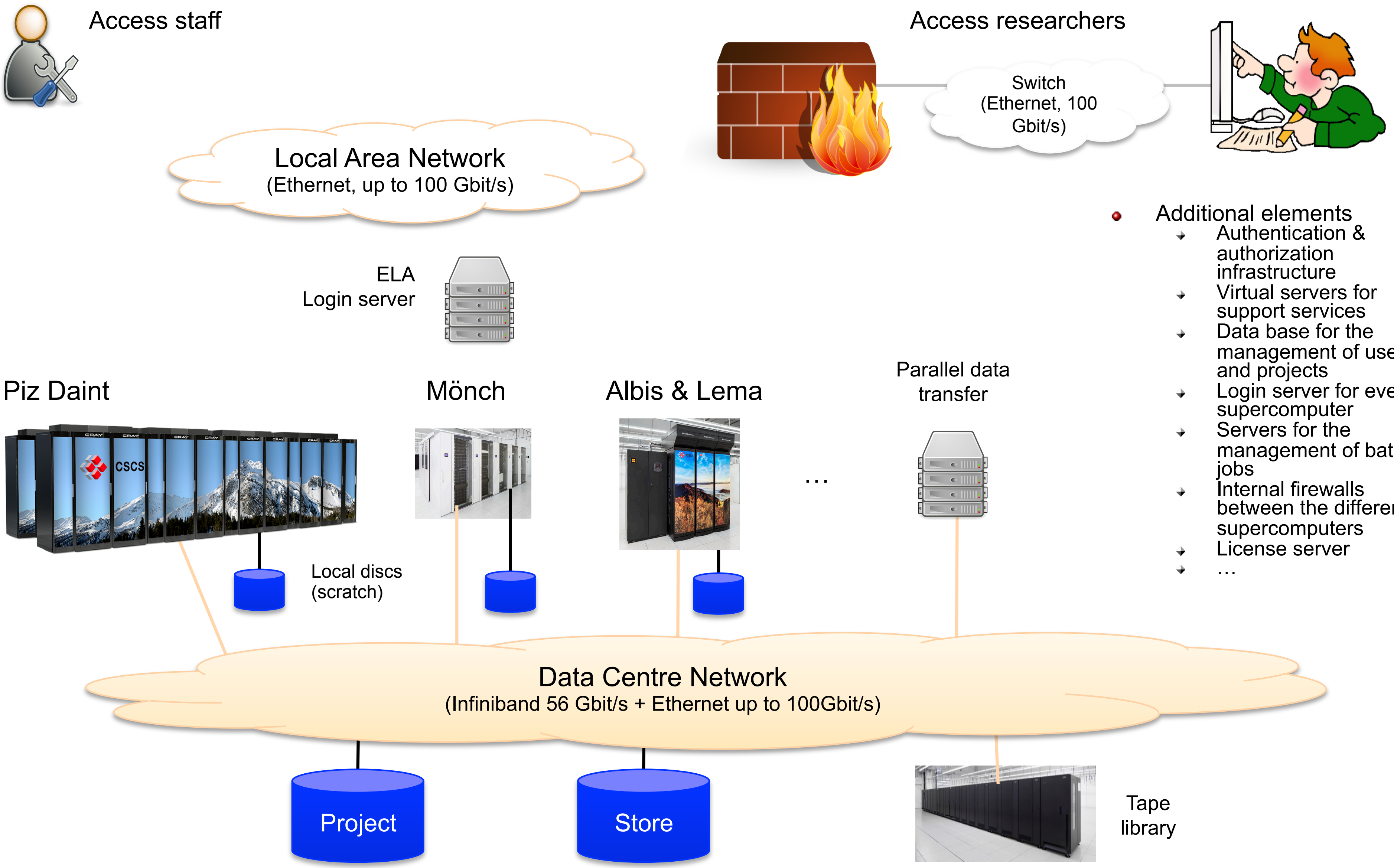
CSCS Systems and Services

- Systems
  - Cray XC30 *28 cabinets*
  - Cray XC40 *7 cabinets*
  - Cray CS-Storm *1 cabinet*
  - Cray CS-Storm *1 cabinet*
  - Cray XE6 *1 cabinet*
  - Cray XE6 *1 cabinet*
  - Commodity *113 nodes*
  - HP *35 nodes*
  - IBM BG/Q *4 racks*
  - NEC *10 racks*

- Services
  - System Management
  - Workload Management (SLURM)
  - Resource Reservations
  - Interactive Environment
  - System Oriented Logging Environment (SOLE)
  - System Level Expertise
  - Customer Specialized
  - Specialized Tools

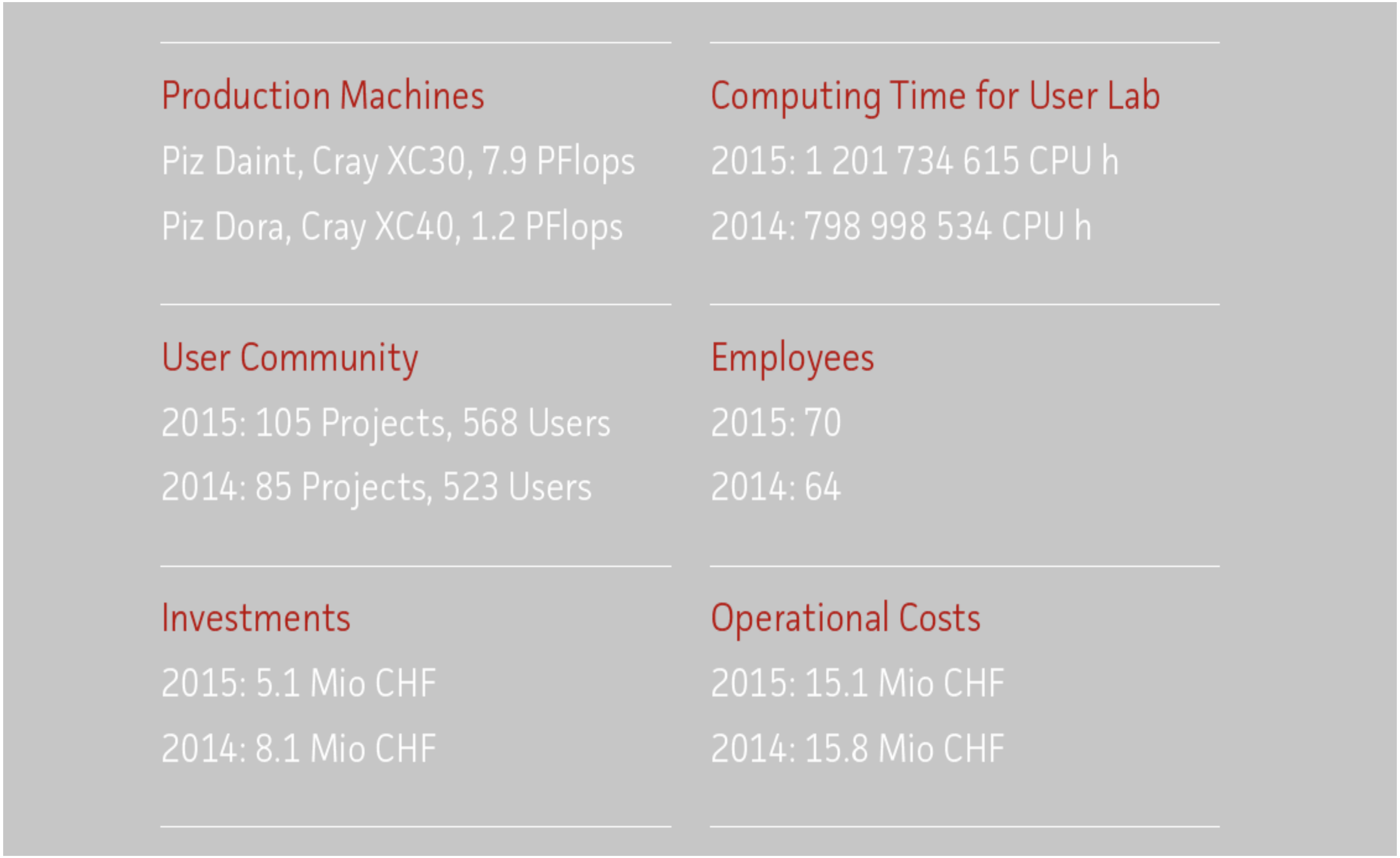
  - Under Investigation
    - Containers
    - Burst Buffer

Overview IT Architecture CSCS



- Additional elements
  - Authentication & authorization infrastructure
  - Virtual servers for support services
  - Data base for the management of users and projects
  - Login server for every supercomputer
  - Servers for the management of batch jobs
  - Internal firewalls between the different supercomputers
  - License server
  - …

CSCS - Key Information



Production Machines

Piz Daint, Cray XC30, 7.9 PFlops

Piz Dora, Cray XC40, 1.2 PFlops

Computing Time for User Lab

2015: 1 201 734 615 CPU h

2014: 798 998 534 CPU h

User Community

2015: 105 Projects, 568 Users

2014: 85 Projects, 523 Users

Employees

2015: 70

2014: 64

Investments

2015: 5.1 Mio CHF

2014: 8.1 Mio CHF

Operational Costs

2015: 15.1 Mio CHF

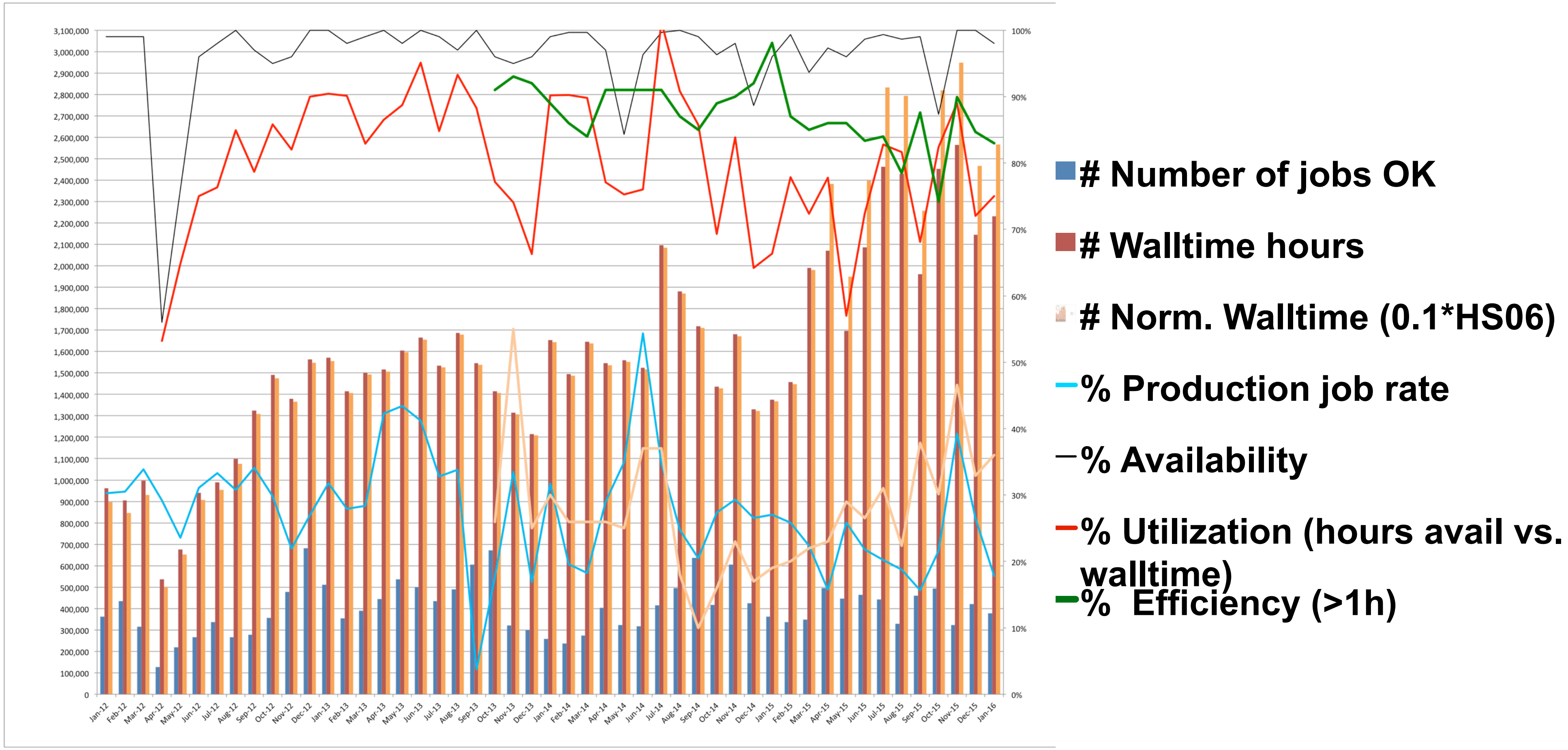
2014: 15.8 Mio CHF

Computing Systems at CSCS

System	Supplier / Model	Installation/ Upgrade	User	Peak Performance (Tflops)
Piz Daint	Cray XC30	2013	User Lab	7787
Piz Dora	Cray XC40	2014	User Lab	1246
Blue Brain 4	IBM BG/Q	2013	EPF Lausanne	839
Piz Kesch & Escha	Cray CS-Storm	2015	MeteoSwiss	305
Mönch	Cluster	2013	ETH Zurich	110
Phoenix	Cluster	2007 / 2012 / 2014	CHIPP (LHC Grid)	65
Albis/Lema	Cray XE6	2012	MeteoSwiss	50
Pilatus	Cluster	2012	User Lab	15

Additional info on T2 info

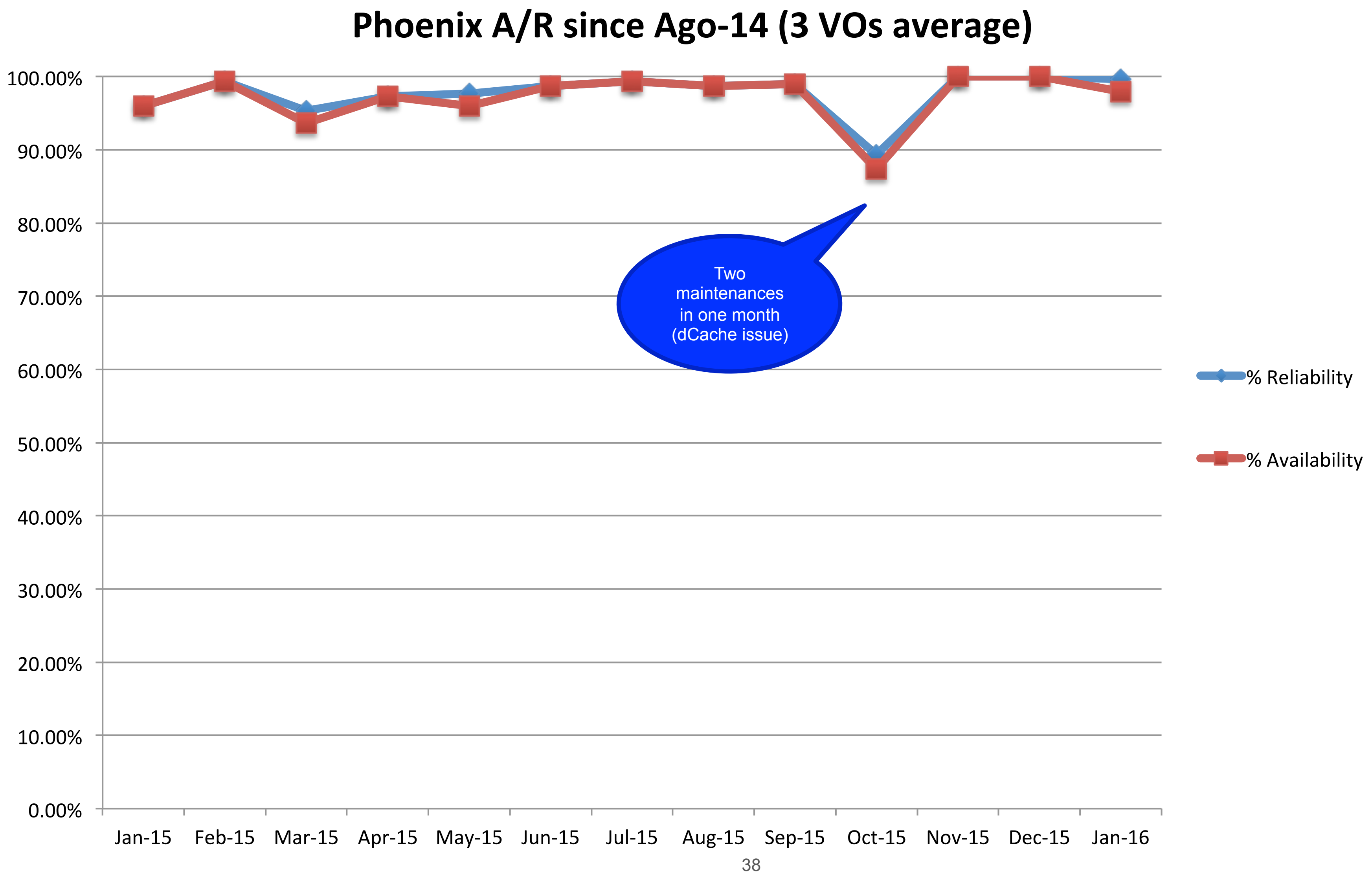
Cluster Performance (1.2012-1.2016)



Overall high availability (>95%) and efficiency typ. 85-95% achieved !
(dip in Apr/May 2012 due to move to Lugano)

- https://wiki.chipp.ch/twiki/bin/view/LCGTier2/WebHome

CSCS Tier-2 – Availability & Reliability


Phoenix A/R since Ago-14 (3 VOs average)
Jan-15
Feb-15
Mar-15
Apr-15
May-15
Jun-15
Jul-15
Aug-15
Sep-15
Oct-15
Nov-15
Dec-15
Jan-16
% Reliability
% Availability
Two maintenances in one month (dCache issue)